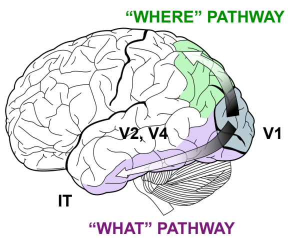
Quantifying the difficulty of object recognition tasks via scaling of accuracy versus training set size



Introduction

Hierarchical models of primate visual cortex (Neo-cognitron/HMAX) have been shown to perform well in object identification tasks [1-5]. We consider the performance of these models as we scale them to the size of human visual cortex, and train them with imagery sets at the scale of human visual experience.

We present quantitative criteria for assessing when a set of learned local representations is complete, based on its statistical evolution with the size of unsupervised learning sets. We also quantify the difficulty of different object recognition tasks via the improvement in classification performance with the size of the supervised training set. Specifically we find a universal form where *accuracy* = *a* + *b* log(*N*), where *a*, and *b* are constants that depend on the details of the system architecture and layer representations and *N* is the number of images in the training set.

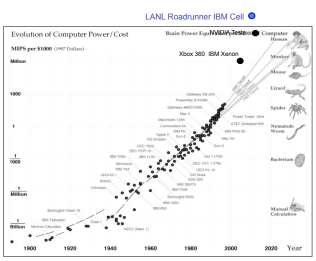


Figure adapted from Hans Moravec, "When will computer hardware match the human brain?", J. Evolution &Technology, 1998.

The Scale of the Human Brain

Units: ~10^{11} neurons
~10^{15} synapses
[~10^4 synapses/neuron]
Temporal rates: 10 Hz
c.f. GHz computer
Performance: 10 PetaFLOPS
c.f. Roadrunner 1.1 Petaflops
Energy consumption: ~20 W
c.f. ~200W GHz computer
Memory: ~synapses 10^{15} bits
c.f. 100 Terabytes computer
Visual Experience: 1 TeraPixel/day,
30 PetaPixels/lifetime

Bounds on Convergence of Classification Accuracy for Large Datasets

Image → x_1 x_2 … x_N
Network → s_1 s_2 … s_N
Classifier → y_1 y_2 … y_N

$$\text{Accuracy}(N) = \frac{p_{1|1}(N) + p_{0|0}(N)}{2} = 1 - \frac{p_{1|0}(N) + p_{0|1}(N)}{2}$$

Binary Classification

The output label y is a variable Bernoulli [p(y|x;α)] such that

$$p_{1|1}(N) = p(y=1|x=1;\alpha^*) \xrightarrow[N\to\infty]{} 1 \qquad p_{0|0}(N) = p(y=0|x=0;\alpha^*) \xrightarrow[N\to\infty]{} 1$$

The task of the classifier is to find the best parameters α* so that the limit of perfect accuracy is approached.

Probably estimation for large data sets

$$p(y_1,y_2,\ldots,y_N|x_1,x_2,\ldots x_N) = \int \prod_{i=1}^{N} p(y_i|x_i;\alpha)\,P(\alpha) = \prod_{i=1}^{N} p(y_i|x_i;\alpha^*) \int \exp(-\ln[\,p(y_i|x_i;\alpha^*)/p(y_i|x_i;\alpha)\,])\ P(\alpha)$$

$$\to \prod_{i=1}^{N} p(y_i|x_i;\alpha^*)\exp[-D_{KL}(\alpha^* \,||\, \alpha)] = [\,p_{1|1}(N)\,]^N$$

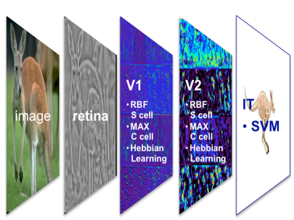
The Kulback –Leibler divergence $D_{KL}(\alpha^* || \alpha)$ measures the 'distance ' between the best estimate α* and the actual estimate after N samples.

For a Bayesian (optimal) estimate of the α in N steps we have (Clark & Barron 1990 [6])

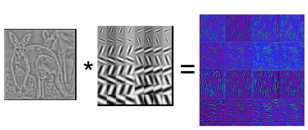
$$D_{KL}(\alpha^* \,||\, \alpha^N) = c/N + (K/2N)\ln N \quad \text{so that } p_{a|b}(N) \rightarrow p^*_{a|b}\,(1 - (K^*/2N)\ln N)\quad a,b=0,1$$

$$\text{Accuracy}(N) = \frac{p^*_{1|1} + p^*_{0|0}}{2} - (K^*/2N)\ln N \quad \text{or} \quad \text{Accuracy}(N) = \frac{p^{ini}_{1|1} + p^{ini}_{0|0}}{2} + (K^{ini}/2N)\ln N$$

Visual Cortex Model



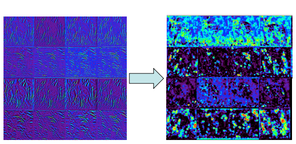
- Neocognitron/HMAX-type hierarchical feed-forward model of visual cortex **"what"** (ventral) pathway (V1 / V2 / IT) with Hebbian learning.
- High performance parallel code using MPI, vector intrinsics, and Cell Broadband Engine.
- Can take as input any image format supported by open source GDAL library, or video format supported by open source FFMPEG library.
- On a cluster of 20 Opteron cores each with a dedicated Cell chip, PANN can process "YouTube"-quality video (200x200 pix) in real time (> 20 fps).



Processing in S cells
Radial Basis Functions with Gabor weight vector.

$$g_{\mathrm{RBF}}(\vec{x},\vec{w}_\Gamma) = \exp\left(-\frac{\beta}{2}(\vec{x}-\vec{w}_\Gamma)^2\right)$$

$$\vec{w}_\Gamma = \vec{w}(\theta,\gamma,\sigma,\lambda,\phi) = \exp\left(-\frac{1}{2\sigma}(x_\theta^2+\gamma y_\theta^2)\right)\cos\left(\frac{2\pi x_\theta}{\lambda}+\phi\right)$$



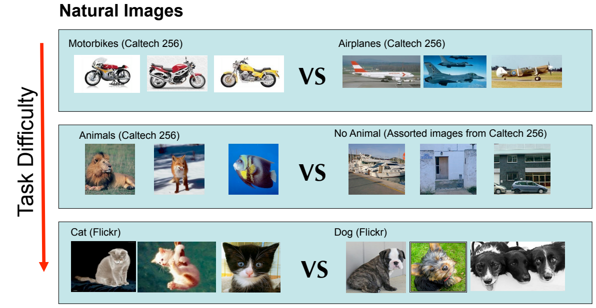
Processing in C cells
MAX Function of S cell receptive fields.

$$g_{\mathrm{MAX}}(\vec{x}) = \max_{i\in I_j}(\{x_i\})$$

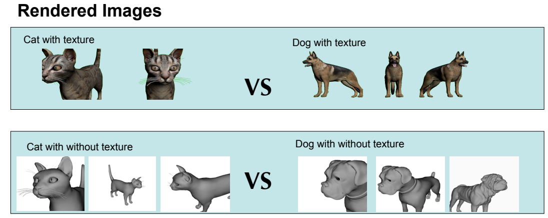
Data Sets

We test our model using standard data sets (Caltech256) [7] and public domain images we selected from Flickr.com. We also consider rendered images using 3ds Max.

Natural Images



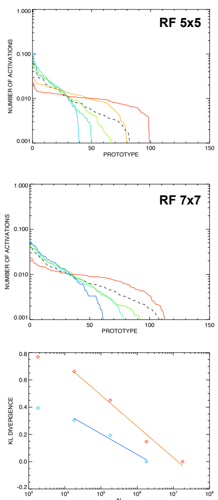
Rendered Images

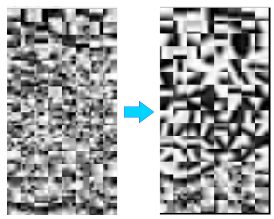


Steven P. Brumby[1], Luís M.A. Bettencourt[1,2], Michael I. Ham[1], Ryan A. Bennett[3], Garrett Kenyon[1]

[1] Los Alamos National Laboratory, Los Alamos NM [2] Santa Fe Institute, Santa Fe NM [3] University of North Texas, Denton TX

Convergence of V1 S-cell Columns for Large Datasets





Hebb - Oha learning rule: $\Delta\vec{w}_j = \alpha \cdot y_j \cdot (\vec{x} - y_j \cdot \vec{w}_j)$

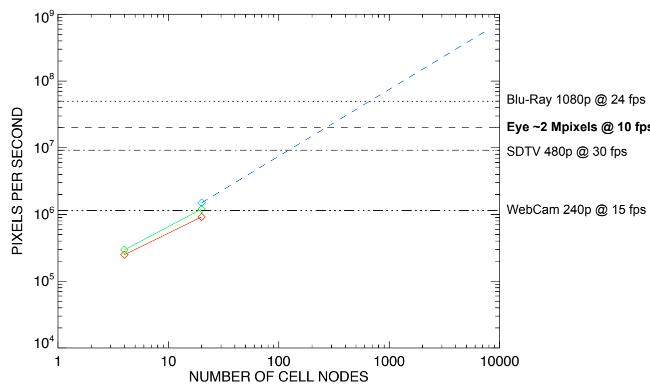
Above: 128 5x5 pixel V1 prototypes sorted by activity after imprinting (*left*) and Hebbian learning (*right*).

Left top: Sorted histogram of number of activations of learned 5x5 prototypes in response to different numbers of retinal patches (cf., grey dashed line = random imprinting): blue=1,800, aqua=18,000, green=180,000, yellow=1,800,000, red = 18,000,000.

Left middle: Sorted histogram of number of activations of learned 7x7 prototypes in response retinal patches (cf., grey dashed line = random imprinting): blue=1,800, aqua=18,000, green=180,000, red = 1,800,000.

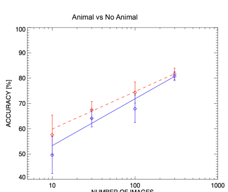
Left bottom: Kullback–Leibler divergence of activation distributions for 5x5 (red points) and 7x7 (blue points), relative to the final (red curves) in upper and middle panels.

Towards full-scale, real-time models of visual cortex



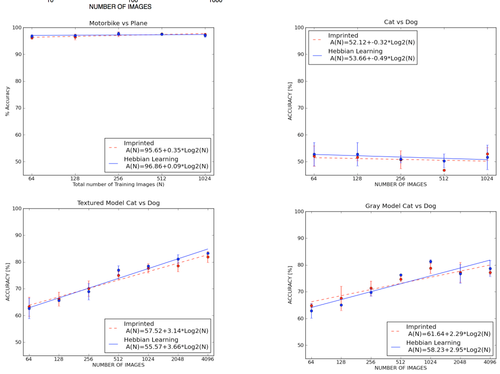
We have built a high-performance implementation of an HMAX-type hierarchical feed-forward model of V1-V2-IT, called PANN [8]. As shown above, on a 20 Opteron core cluster, where each core has access to a dedicated Cell Broadband Engine, we currently reach processing levels above 1M pixels/second, sufficient for real-time processing of webcam-quality video streams. LANL's petascale computing machine, Roadrunner, consists of ~10,000 Cell -accelerated cores, so that even with less than ideal (linear) scaling (blue dashed line above), we expect to process human eye-like video streams in real-time.

Scaling of IT Classifier Performance for Large Datasets



IT is modeled by a conventional binary classifier, typically a support vector machine (SVM). We show performance of the IT classifier for the datasets introduced previously.

In each case, we show performance for the standard V2 imprinting algorithm (Serre, et al., 2007 [4]) (red dashed line) and for a V2 whose tunings are set using Hebbian learning (blue solid line).



Conclusions

- Why is the visual system so large? To match the amount of visual experience? Can large-scale models approach human performance?
- The brain has a (very large) finite number of parameters that are learned through visual experience, and there are universal bounds on how fast a finite system (however large) can learn.
- More complex object classes require in general more parameters for the same accuracy and a commensurate amount of visual experience (N > K).
- The universal bounds correspond to optimal learning from examples and control both the (unsupervised) learning of neuronal tunings (in V1 and other layers) and the accuracy of object recognition.

References

[1] D.H. Hubel, T.N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex”, J. Physiol. (Lond.) 195, 215–243 (1968).
[2] K. Fukushima: Neocognitron, “A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, Biological Cybernetics, 36(4), pp. 193-202, April 1980.
[3] Riesenhuber, M. & Poggio, T., “Hierarchical Models of Object Recognition in Cortex”, Nature Neuroscience 2: 1019-1025, 1999.
[4] T. Serre, A. Oliva and T. Poggio. “A feedforward architecture accounts for rapid categorization”, Proceedings of the National Academy of Science, 104 (15), pp. 6424-6429, April 2007.
[5] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, “Object recognition with cortex-like mechanisms”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (3), pp. 411-426, 2007.
[6] B.S. Clarke and A.R. Barron, “Information-Theoretic Asymptotics of Bayes Methods”, IEEE Trans. on Information Theory, 36 (3), pp. 453-471, May 1990.
[7] L. Fei-Fei, R. Fergus and P. Perona, “Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories”, IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004.
[8] Steven P. Brumby, Garrett Kenyon, Will Landecker, Craig Rasmussen, Sriram Swaminarayan, and Luís M. A. Bettencourt, “Large-scale functional models of visual cortex for remote sensing”, 2009 38th IEEE Applied Imagery Pattern Recognition, Vision: Humans, Animals, and Machines, Cosmos Club, Washington DC October 14-16, 2009